ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores

Sandeep Mathias, Pushpak Bhattacharyya

Centre for Indian Language Technology
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
{sam, pb}@cse.iitb.ac.in

Introduction

Motivation

- Since 2012, the *de facto* dataset for automatic essay grading (AEG) has been the Automated Student's Assessment Prize (ASAP) AEG dataset.
- The ASAP AEG dataset has 8 topics, with about 13,000 essays. However, only 2 prompts have scores for different essay attributes.

Contribution

Creation of a resource for essay attribute scores.

Related Work

- While there has been a lot of work done in overall essay scoring, not much has been done with respect to scoring particular attributes of essays.
- Some of the attributes that have been scored include organization (Persing et al., 2010), prompt adherence (Persing and Ng, 2014), coherence (Somasundaran et al., 2014).

Dataset

Essays

The entire ASAP dataset has nearly 13,000 essays across 8 prompts. 6 of those 8 prompts, comprising nearly 10,400 essays, only have an overall score.

Prompt ID	Essay Type	Essays	Length	Scores	Attribute Scores
Prompt 1	Argumentative	1785	350	1 - 6	**No**
Prompt 2	Argumentative	1800	350	1 - 6	**No**
Prompt 3	Source-Dependent	1726	150	0 - 3	**No**
Prompt 4	Source-Dependent	1772	150	0 - 3	**No**
Prompt 5	Source-Dependent	1805	150	0 - 4	**No**
Prompt 6	Source-Dependent	1800	150	0 - 4	**No**
Prompt 7	Narrative	1569	300	0 - 3	**Yes**
Prompt 8	Narrative	723	650	1 - 6	**Yes**

Table 1: *Description of the ASAP AEG dataset.*

Table 1 gives a description of the different essay prompts. Since scores are already present for prompts 7 & 8, we mainly provide scores for prompts 1 to 6.

Types of Essays

There are 3 types of essays in the dataset.

1. **Argumentative / Persuasive essays** - These are essays where the prompt is one in which the writer has to convince the reader about their stance for or against a topic (for example, free speech in public colleges).
2. **Source-dependent responses** - These essays are responses to a source text, where the writer responds to a question about the text (for instance, describing the writer's opinion about an incident that happened to him in the text).
3. **Narrative / Descriptive essays** - These are essays where the prompt requires us to describe / narrate a story.

Attributes of Essays

Based on the types of essays, there are 2 sets of attributes.

Attributes of Argumentative / Persuasive Essays

There are 5 attributes for narrative essays, namely

1. **Content**: The quantity of relevant text present in the essay.
2. **Organization**: The way the essay is structured.
3. **Word Choice**: The choice and aptness of the vocabulary used in the essay.
4. **Sentence Fluency**: The quality of the sentences in the essay.
5. **Conventions**: Overall writing conventions to be followed, like spelling, punctuations, etc.

Attributes of Source-dependent Responses

There are 4 attributes for source-dependent responses, namely

1. **Content**: The amount of relevant text present in the essay.
2. **Prompt Adherence**: A measure of how the writer sticks to the question asked in the prompt.
3. **Language**: The quality of the grammar and spelling in the response.
4. **Narrativity**: A measure of the coherence and cohesion of the response to the prompt.

Creation of the Resource

- Each essay in a particular prompt was score by an annotator.
- In case an essay's attribute scores differed by a large margin from the original overall score, a second annotator scored the essay.

Annotator Details

- The entire dataset was scored by 3 annotators.
- All the annotators either had a Master's Degree in English or were currently studying for one.

Baseline Experiments

Feature Type	Feature List
Length	Word Count, Sentence Count, Sentence Length, Word Length
Punctuation	Counts of Commas, Quotations, Apostrophes, etc.
Syntax	Parse Tree Depth, Subordinate Clauses, etc.
Stylistic Features	Formality, Word Frequency, Type-Token Ratio
Cohesion	Discourse Connectives, Entity Grid, etc.
Coherence	Average Similarity between adjacent sentences of PoS tags, Lemmas, etc.
Language Model	Count of OOVs, LM score, etc.
n-Grams	Word n-Grams and PoS n-Grams

Table 2: *Different features used in our experiment (Zesch et al., 2015)*

Classifier used: Random Forest
Evaluation Metric: Cohen's Kappa with Quadratic Weights (Cohen, 1968)
Features used: Task-independent features from Zesch et al. (2015).
Evaluation Method: 5-fold cross-validation.

Results

Prompt ID	Cont.	Org.	WC	SF	Conv.	Overall
Prompt 1	0.67	0.60	0.64	0.62	0.61	0.74
Prompt 2	0.61	0.58	0.60	0.59	0.62	0.62

Table 3: *Results of the 5-fold cross-validation for argumentative / persuasive essays.*

Prompt ID	Cont.	PA	Lang.	Narr.	Overall
Prompt 3	0.59	0.59	0.57	0.63	0.54
Prompt 4	0.66	0.66	0.56	0.67	0.68
Prompt 5	0.67	0.64	0.60	0.63	0.76
Prompt 6	0.60	0.56	0.59	0.62	0.63

Table 4: *Results of the 5-fold cross-validation results using the Random Forest classifier for source-dependent essays.*

Prompt ID	Content	Organization	Word Choice	Sentence Fluency	Conventions
Prompt 1	Coherence	Length	Coherence	Syntax	Coherence
Prompt 2	Coherence	Coherence	Coherence	Syntax	Coherence
Average	Coherence	Coherence	Coherence	Syntax	Coherence

Table 5: *Results of the ablation tests using the Random Forest classifier for argumentative / persuasive essays to determine the most important feature set for each attribute in each prompt.*

Prompt ID	Content	Prompt Adherence	Language	Narrativity
Prompt 3	Length	Coherence	Coherence	Style
Prompt 4	Punctuation	Language Model	Coherence	Complexity
Prompt 5	Length	Coherence	Punctuation	Language Model
Prompt 6	Coherence	Language Model	Coherence	Coherence
Average	Length	Coherence	Coherence	Coherence

Table 6: *Results of the ablation tests using the Random Forest classifier for source-dependent responses to determine the most important feature set for each attribute in each prompt*

Conclusion

- We present a manually annotated dataset for automated essay grading.
- Annotations can be used as a gold standard for future experiments involving attribute-specific scoring of essays.
- The resource is available online at **`https://cfilt.iitb.ac.in/~egdata/`**

References

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Persing, I., Davis, A., and Ng, V. (2010). Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA. Association for Computational Linguistics.

Persing, I. and Ng, V. (2014). Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, Maryland. Association for Computational Linguistics.

Somasundaran, S., Burstein, J., and Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 950–961, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Zesch, T., Wojatzki, M., and Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–232, Denver, Colorado. Association for Computational Linguistics.

Presented at the 11th edition of the Language Resources and Evaluation Conference, 7-12 May 2018, Miyazaki (Japan)